Supplemental Information

Sequences Associated with Centromere Competency in the Human Genome

Karen E. Hayden[1,2$], Erin D. Strome[1,3$], Stephanie E. Merrett[1], Hye-Ran Lee[1,4], M. Katharine Rudd[1,5] and Huntington F. Willard[1]*

[1] Genome Biology Group, Duke Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina, United States of America

[$] Authors contributed equally to this work

[2] Present address: Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA, United States of America

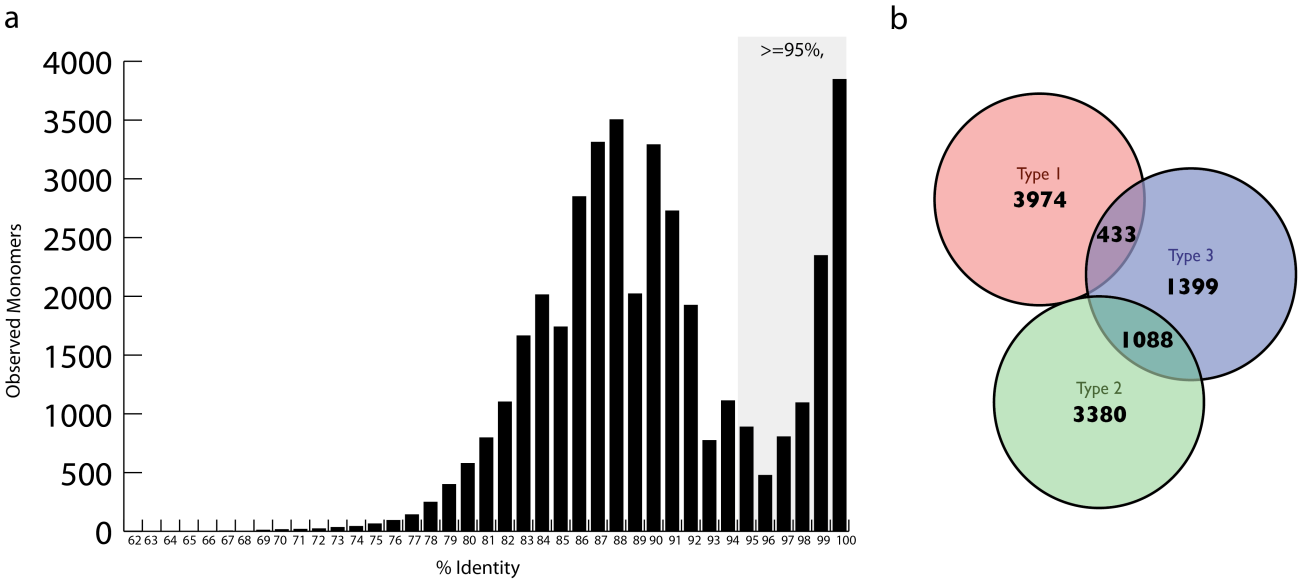
[3] Present address: Northern Kentucky University, Highland Heights, KY, United States of America

[4] Present address: University of North Carolina, Chapel Hill, NC, United States of America

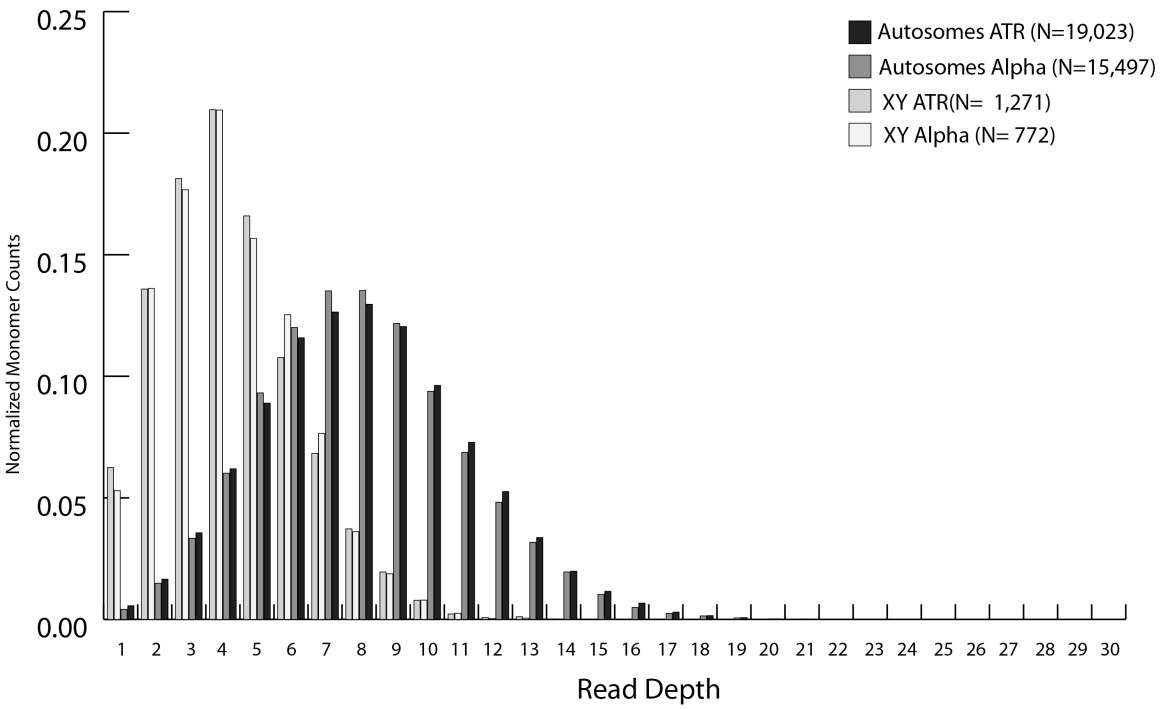
[5] Present address: Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, United States of America

*Address for Correspondence:
Dr. Huntington F. Willard
Duke Institute for Genome Sciences & Policy
Duke University
CIEMAS Room 2379
101 Science Drive
Durham, NC, 27708
Phone: 919-668-4477
e-mail: hunt.willard@duke.edu

Supplemental Data

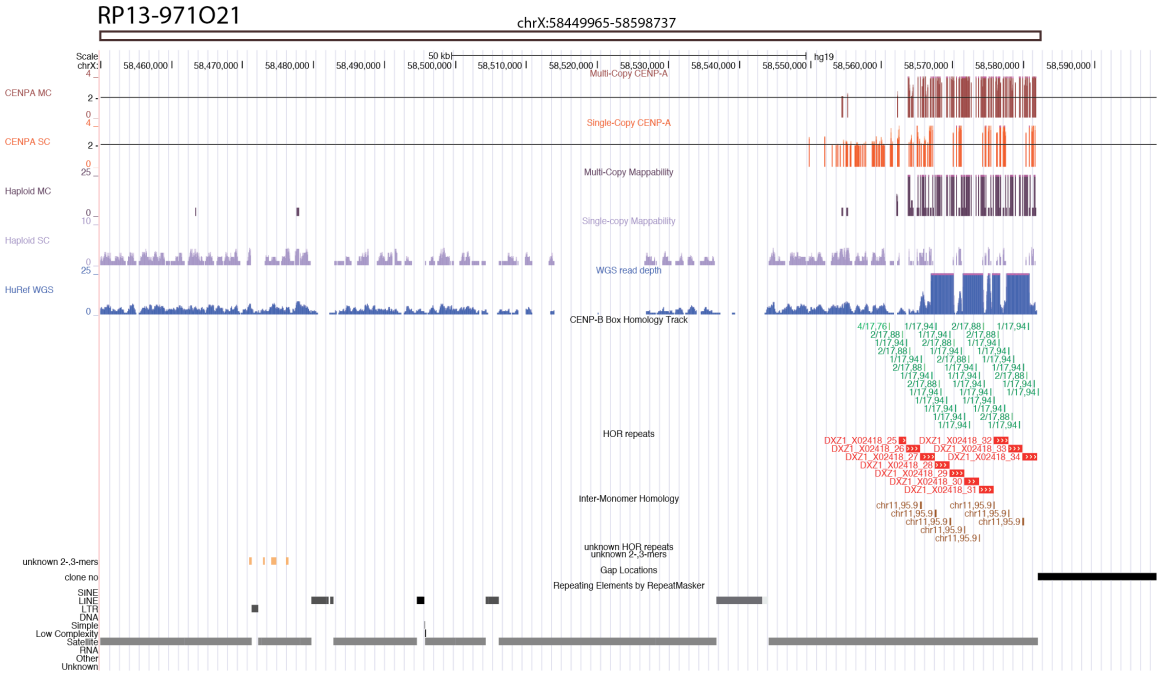


Supplemental Figure 1. Characterization of monomer homology types in the human reference assembly. (a) Distribution of top non-self percent identity of pairwise alignments for each full-length monomer in the human reference assembly. Grey highlighting indicates the threshold of 95% identity used throughout this study. (b) Among the monomers that pass the 95% threshold, three types of monomer homology patterns were detected: tandem monomer patterns often characterized as higher-order repeats (type 1); those monomers that are not organized in tandem, yet share high homology with monomer or monomers within the same chromosome (type 2); or those monomers that are not organized in tandem, yet share high homology with monomer or monomers on different chromosomes (type 3). These monomer types are not exclusive, as we observe monomers that are involved in both intra- and inter-chromosomal relationships, as shown.

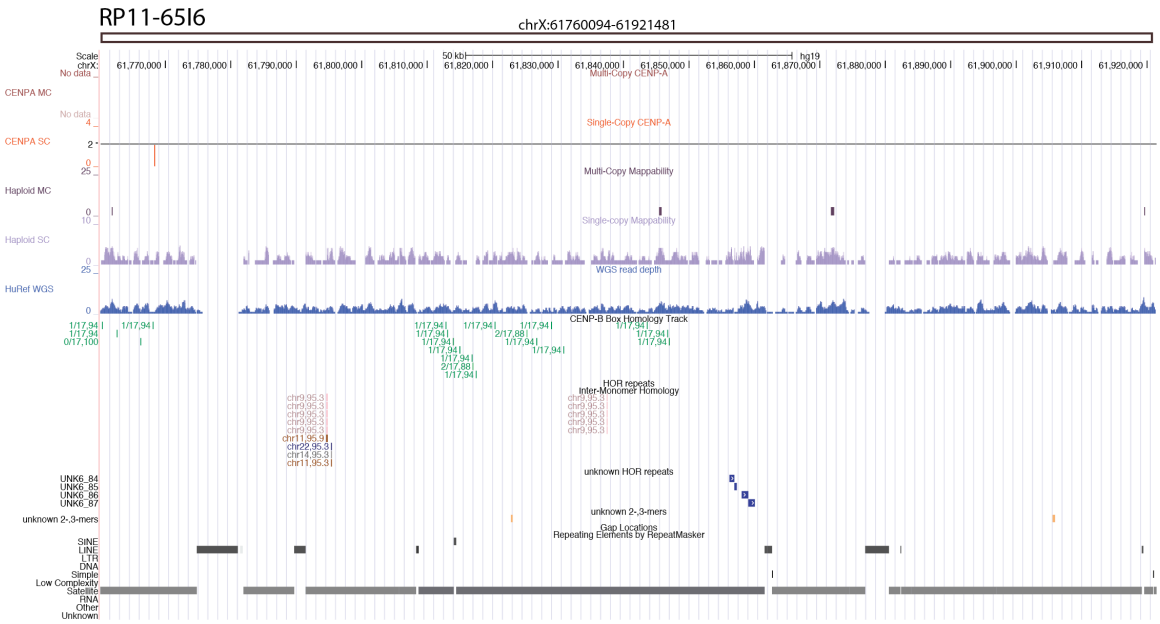


Supplemental Figure 2. Read depth estimates for single-copy, divergent alpha satellite DNA relative to non-repeat AT-rich sites in the reference genome (GRch37/hg19). Read depth estimates of HuRef read alignments in assembled regions containing monomers from non-homogenized repeats (Alpha) or non-repeat AT-rich (ATR) regions demonstrate that alpha satellite read depth patterns are not subject to sequencing bias and are defined by the same sequence coverage as known euchromatic single-copy DNA. Unique genomic sites in the human genome were identified by the lack of annotated repeats (RepeatMasker, tandem repeatFinder, and segmental duplications/copy number variation) and by AT-richness corresponding to the observed AT frequency in alpha satellite DNA (≥ 58% AT). A total of 20,294 single-copy 171-bp windows (representing a control database of non-alpha single copy sequences) were selected from these unique sites (19,023 from autosomes and 1,271 from X and Y chromosomes). HuRef WGS read alignment estimates reflect the estimated coverage for each base in the unique genomic database and those alpha satellite sequences that are characterized as single-copy (representing 15,497 monomers from autosomal pericentromeric regions and 772 monomers from the X and Y chromosome).

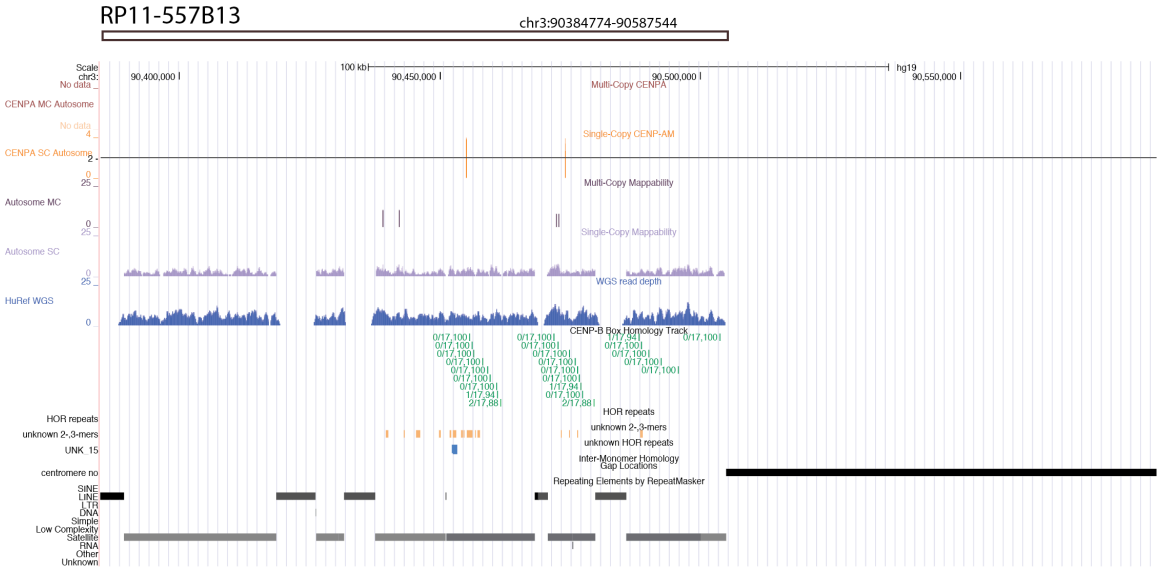
a


RP13-971O21
chrX:58449965-58598737
Multi-Copy CENP-A
Single-Copy CENP-A
Multi-Copy Mappability
Single-copy Mappability
WGS read depth
CENP-B Box Homology Track
HOR repeats
Inter-Monomer Homology
unknown HOR repeats
unknown 2-,3-mers
Gap Locations
Repeating Elements by RepeatMasker
CENPA MC
CENPA SC
Haploid MC
Haploid SC
HuRef WGS
clone no
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

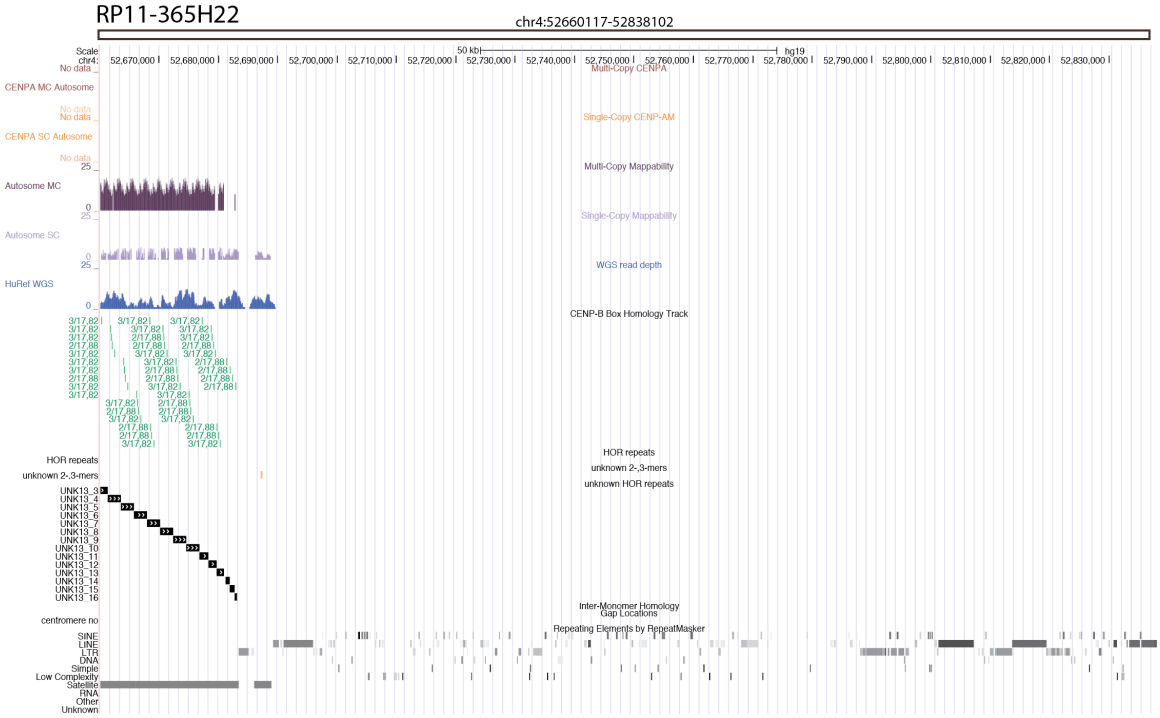
b


RP11-65I6
chrX:61760094-61921481
Multi-Copy CENP-A
Single-Copy CENP-A
Multi-Copy Mappability
Single-copy Mappability
WGS read depth
CENP-B Box Homology Track
HOR repeats
Inter-Monomer Homology
unknown HOR repeats
unknown 2-,3-mers
Gap Locations
Repeating Elements by RepeatMasker
CENPA MC
CENPA SC
Haploid MC
Haploid SC
HuRef WGS
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

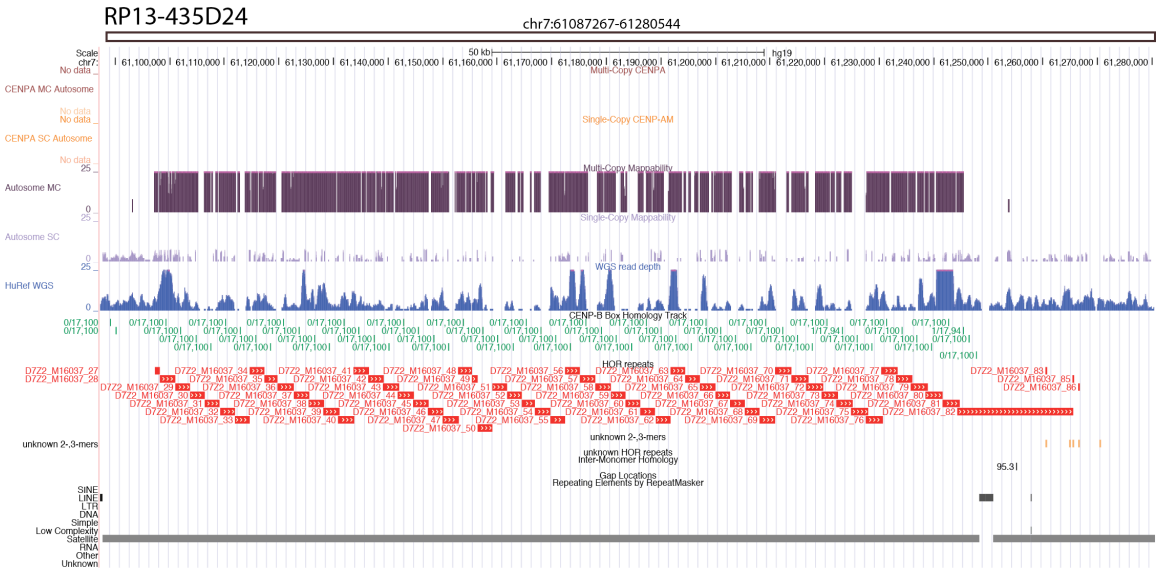
c


RP11-557B13
chr3:90384774-90587544
Scale
chr3:
100 kb
hg19
90,400,000
90,450,000
90,500,000
90,550,000
No data
Multi-Copy CENPA
CENPA MC Autosome
No data
Single-Copy CENP-AM
CENPA SC Autosome
Multi-Copy Mappability
Autosome MC
Single-Copy Mappability
Autosome SC
WGS read depth
HuRef WGS
CENP-B Box Homology Track
HOR repeats
unknown 2-,3-mers
UNK_15
unknown HOR repeats
Inter-Monomer Homology
centromere no
Gap Locations
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

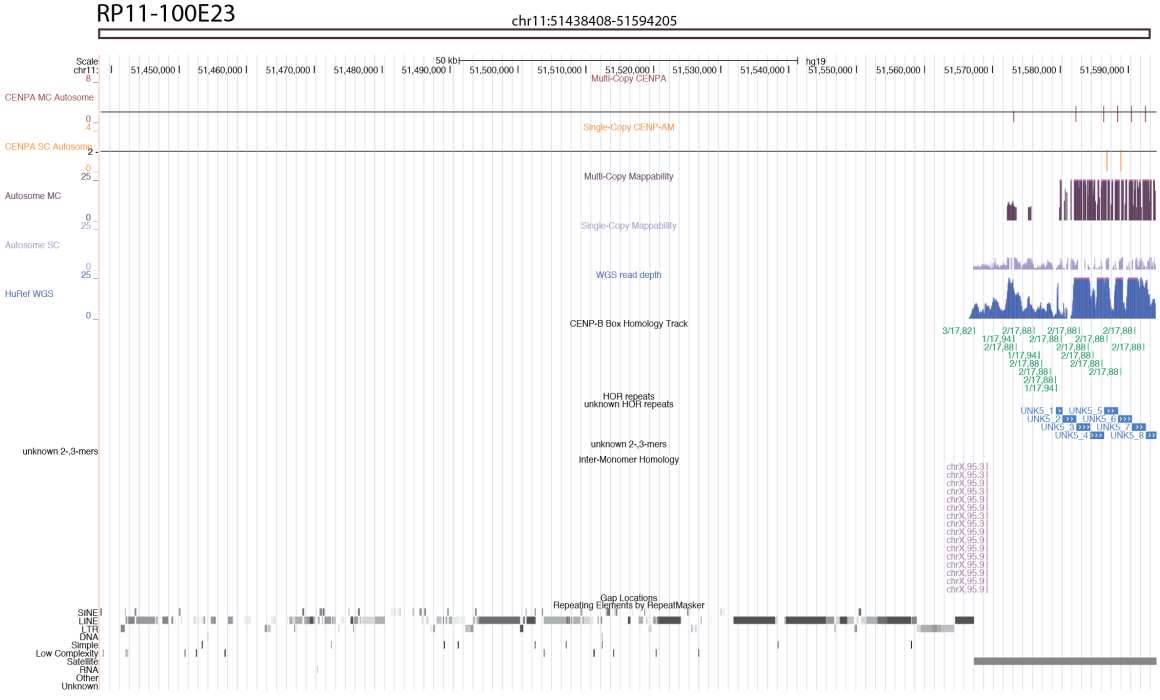
d


RP11-365H22
chr4:52660117-52838102
Scale
chr4:
50 kb
hg19
52,670,000
52,680,000
52,690,000
52,700,000
52,710,000
52,720,000
52,730,000
52,740,000
52,750,000
52,760,000
52,770,000
52,780,000
52,790,000
52,800,000
52,810,000
52,820,000
52,830,000
No data
Multi-Copy CENPA
CENPA MC Autosome
No data
Single-Copy CENP-AM
CENPA SC Autosome
Multi-Copy Mappability
Autosome MC
Single-Copy Mappability
Autosome SC
WGS read depth
HuRef WGS
CENP-B Box Homology Track
HOR repeats
unknown 2-,3-mers
unknown HOR repeats
UNK13_3
UNK13_4
UNK13_5
UNK13_6
UNK13_7
UNK13_8
UNK13_9
UNK13_10
UNK13_11
UNK13_12
UNK13_13
UNK13_14
UNK13_15
UNK13_16
Inter-Monomer Homology
centromere no
Gap Locations
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

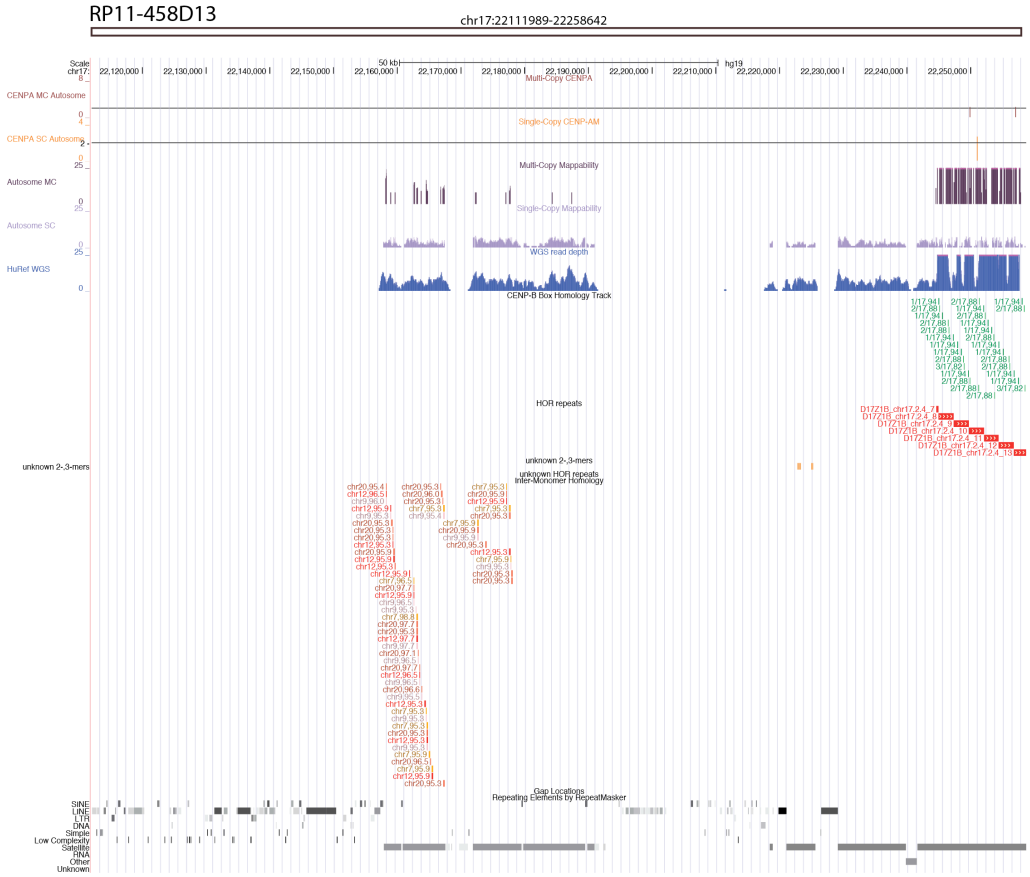
e


RP13-435D24
chr7:61087267-61280544
Multi-Copy CENPA
Single-Copy CENP-AM
Multi-Copy Mappability
Single-Copy Mappability
WGS read depth
CENP-B Box Homology Track
HOR repeats
unknown 2-,3-mers
unknown HOR repeats
Inter-Monomer Homology
Gap Locations
Repeating Elements by RepeatMasker

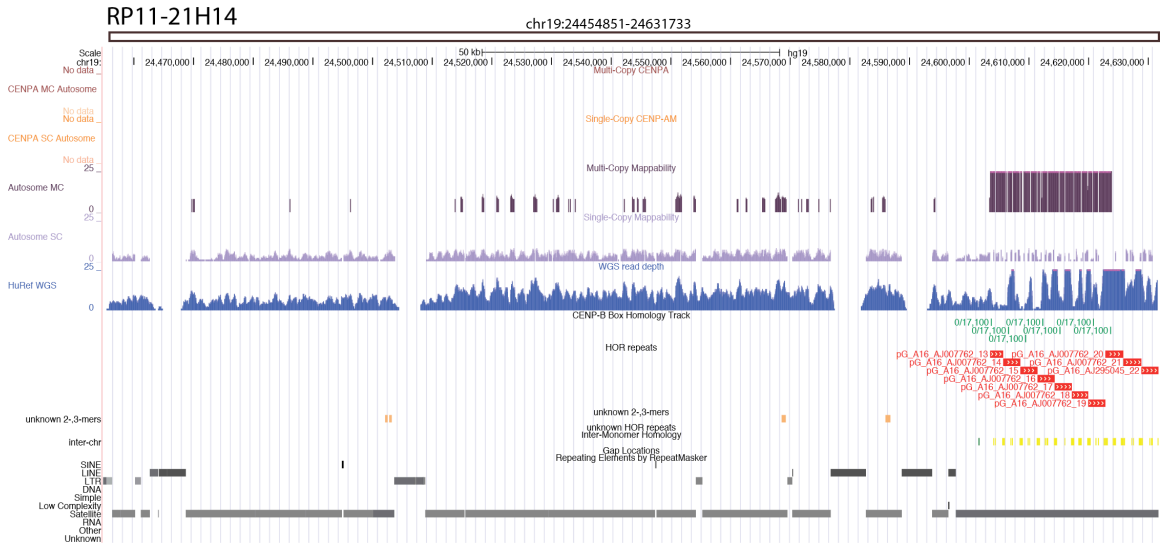
f


RP11-100E23
chr11:51438408-51594205
Multi-Copy CENPA
Single-Copy CENP-AM
Multi-Copy Mappability
Single-Copy Mappability
WGS read depth
CENP-B Box Homology Track
HOR repeats
unknown HOR repeats
unknown 2-,3-mers
Inter-Monomer Homology
Gap Locations
Repeating Elements by RepeatMasker

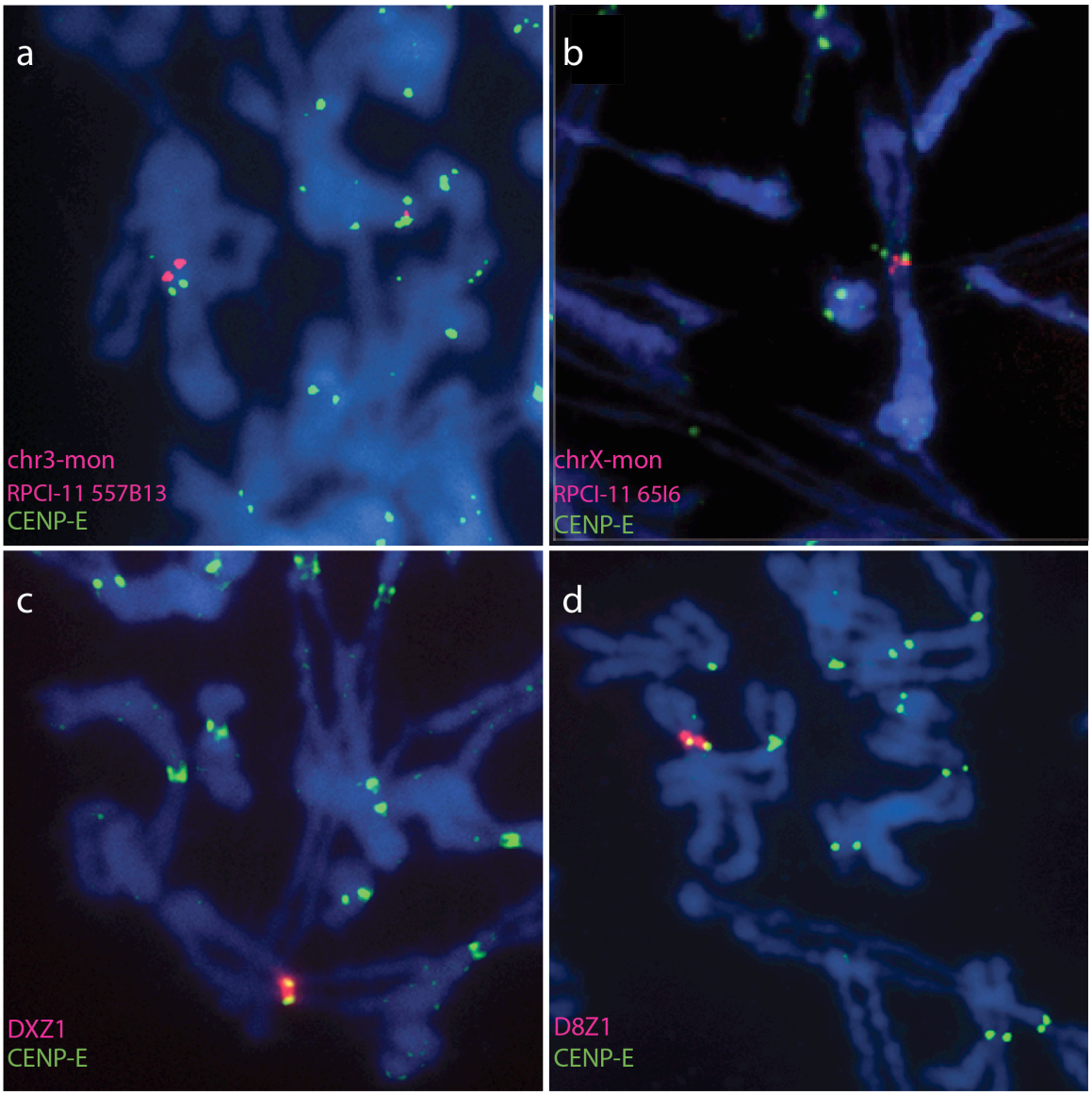
g


RP11-458D13
chr17:22111989-22258642
Multi-Copy CENPA
Single-Copy CENP-AM
Multi-Copy Mappability
Single-Copy Mappability
WGS read depth
CENP-B Box Homology Track
HOR repeats
unknown 2-,3-mers
unknown HOR repeats
Inter-Monomer Homology
Gap Locations
Repeating Elements by RepeatMasker
CENPA MC Autosome
CENPA SC Autosome
Autosome MC
Autosome SC
HuRef WGS
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

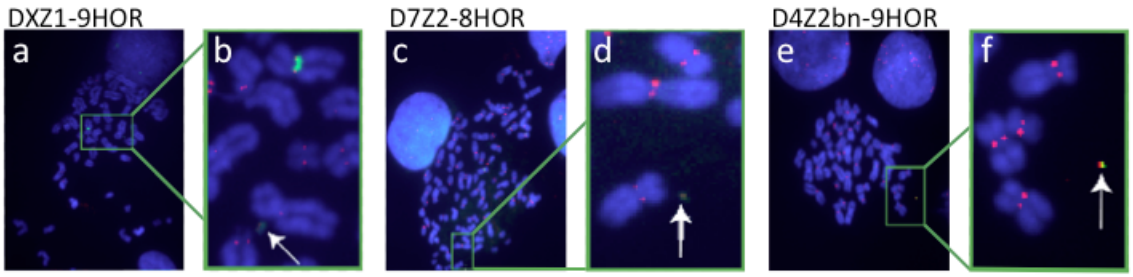
h


RP11-21H14
chr19:24454851-24631733
Multi-Copy CENPA
Single-Copy CENP-AM
Multi-Copy Mappability
Single-Copy Mappability
WGS read depth
CENP-B Box Homology Track
HOR repeats
unknown 2-,3-mers
unknown HOR repeats
Inter-Monomer Homology
Gap Locations
Repeating Elements by RepeatMasker
CENPA MC Autosome
CENPA SC Autosome
Autosome MC
Autosome SC
HuRef WGS
inter-chr
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Supplemental Figure 3a-h: GRCh37/hg19 annotation maps for BACs tested in human artificial chromosome assays. Alpha satellite annotations are presented (a-h) for each BAC and corresponding genomic region (GRCh37/hg19). Identification for each BAC and genomic coordinates are along the top of each UCSC browser image (Kent et al. 1997). Monomer homology patterns are provided as individual tracks: intra-chromosomal (type 1) monomer patterns are provided as known higher-order repeats (HOR) (red), unknown HOR (blue), and 2- and 3-mers (yellow); inter-chromosomal homologies (type 3) are listed with chromosomal color key and label provided. CENP-B boxes are indicated in green, with labels indicating the distance from the alpha satellite consensus CENP-B motif (CTTCGTTGGAAACGGG). Read depth estimates are provided in blue, with corresponding 50-mer mappability tracks subdivided into single-copy (light purple) and multi-copy sites (dark purple). Those sites of CENP-A enrichment, defined as ≥ 2 log transformed ratio of relative frequencies of 50-mers of IP /Mock, are indicated as either single-copy (light orange) or multi-copy (dark orange) corresponding to the mappability track below. For example, (a) DXZ1 is enriched for CENP-A, as observed across both single-copy and multi-copy mappability tracks. With exception of those few sites that show CENP-A enrichment (e.g. (b) chromosome X monomeric), the BAC regions selected in the study appear to generally lack CENP-A enrichment, suggesting that they are non-functioning at the genomic level in the HuRef cell line.



Supplemental Figure 4. BACs containing monomeric alpha satellite do not co-localize with centromeric chromatin in HT1080 cells. Extended chromosomes (see Supplemental Methods) from HT1080 cells demonstrate the presence or absence of centromeric chromatin (defined by the presence of CENP-E, in green). BACs containing monomeric alpha satellite used in this study: (a) RP11-557B13 and (b) RP11-65I6, shown in red, do not overlap with CENP-E, as observed for DXZ1 and D8Z1 higher-order repeat arrays (c and d).



Supplemental Figure 5. Representative results of human artificial chromosome assays. Human artificial chromosomes derived from BACs from Xp:DXZ1-9HOR (RP11-971O21,a-b), 7q:D7Z2-8HOR (RP11-435D24, c-d), and 4q:D4Z2bn-9HOR (RP11-365H22, e-f) are identified by FISH (using HOR-specific alpha-satellite DNA, shown in green) and shown to co-localize with centromere chromatin, centromere protein A (CENP-A, shown in red).

Supplemental Table Legends

Supplemental Table 1. Type 1 alpha satellite monomer homology patterns detected in GRCh37/hg19. Column header information is defined as follows: chr, hg19 chromosome; chrS, chromosome start position of monomer; chrE, chromosome end position of monomer; bp_span, the length of the repeat unit (chrE-chrS); HOR, the name of the characterized array referenced from available literature; HOR_lit, the corresponding citation for HOR provided.

Supplemental Table 2. Type 2 alpha satellite monomer homology patterns detected in GRCh37/hg19. Column header information is defined as follows: chr, hg19 chromosome; chrS, chromosome start position of monomer; chrE, chromosome end position of monomer; chrS_hom, chromosome start of intra-chromosomal monomer homology patterns; chrE_hom, chromosome end of intra-chromosomal monomer homology patterns; monPID, monomer percent identity (provided if ≥ 95%); distance, bp distance between monomers (3' monomer start - 5' monomer end).

Supplemental Table 3. Type 3 alpha satellite monomer homology patterns detected in GRCh37/hg19. Column header information is defined as follows: chr, hg19

chromosome; chrS, chromosome start position of monomer; chrE, chromosome end position of monomer; chr_hom, hg19 chromosome of monomer homology; chrS_hom, chromosome start of inter-chromosomal monomer homology patterns; chrE_hom, chromosome end of inter-chromosomal monomer homology patterns; monPID, monomer percent identity (provided if ≥ 95%); SegDup_overlap, number of bases of overlap with segmental duplication UCSC annotation track.

Supplemental Table 4. CENP-B box annotation from GRCh37/hg19 assembly. Column header information is defined as follows: chr, hg19 chromosome; chrS, chromosome start position of CENP-B box motif; chrE, chromosome end position of CENP-B box motif; CENP-B_diff, number of bases that differ from consensus sequence (CTTCGTTGGAAACGGG); CENP-B_motif, CENP-B box motif sequence identified.

Supplemental Table 5. Summary of chromosomal mappable sites and CENP-A enrichment. Results are provided for each reference chromosome assembly with respect to mappability and CENP-A enrichment. Mappability estimates are subdivided by their association with higher-order repeats (HOR) (including dimers and trimers, often in higher abundance in the reference genome), non-HOR monomers (intra- and inter-), and those monomers that are characterized as monomeric. Corresponding sites (50-mer) of CENP-A enrichment (≥2 fold) are provided for each mappability category.

Supplemental Table 6. CENP-A enrichment summary within assembled higher-order repeat sequences. Higher-order repeats (HOR_ID) on each reference chromosome assembly (chr) are listed with respect to the number of sites determined to be mappable (Mappable Sites), and the number of sites that are observed to be enriched by CENP-A (CENP-A Enriched Sites), with the percentage of CENP-A enriched within total mappable sites provided (% Sites Enriched).

Supplemental Table 7. CENP-A enrichment detected within sites of low/single coverage, monomeric alpha satellite DNA. All regions reported must have two or more overlapping 50-mers with CENP-A enrichment (≥2 fold). Column header information is defined as follows: UID (unique identifier), number of each catalogued site, chr, hg19

chromosome; chrS, chromosome start position series of overlapping enriched 50-mers; chrE, chromosome end position of series of overlapping enriched 50-mers; Number of overlapping 50-mers, the number of 50mer windows between chrS and chrE.

Supplemental Experimental Procedures

Extended Chromosome Protocol

General chromosome slide preparation as previously described [1]. Highly confluent HT1080 cells (90-95% confluency) were treated with ethidium bromide to a final concentration of 1.25×10^{-5} M for one hour prior to adding colcemid (150μg) for 30 minutes. Cells were placed in hypotonic solution (0.075M KCl) for 12 min at 37°C before cytospin treatment (Shandon Cytospin 3) for 10 min at 2000 RPM. Fluorescence *in situ* hybridization and immunostaining procedures were performed as described [2, 3], using a polyclonal rabbit CENP-E antibody.

Alpha Satellite Homology Patterns in the Reference Human Genome

To characterize alpha satellite sequence content and organization, we performed pairwise alignments to assess patterns of intra- and/or inter-chromosomal sequence homology with 39,986 full-length monomers obtained from the current human genome (hg19), representing 6.8 Mb of assembled alpha satellite sequence assigned to individual chromosomes. Approximately three-quarters of the monomers (29,712) showed no evidence of high identity to any other sequences in the assembly and thus could be annotated as divergent monomeric sequences [4-6]. The remaining 10,274 monomers, however, showed signatures of recent satellite expansion and/or homogenization. Three patterns of homogenization were detected: type 1: local homogenization, where either single monomers or multimeric blocks of monomers have at least one near-identical tandem repeat unit (4407 monomers, 42.9% of all homogenized monomers); type 2: monomers that share high *intra*-chromosomal identity, but that are not in tandem (4468 monomers, 43.4% of all homogenized monomers); and type 3: monomers that share high *inter*-chromosomal identity (2920 monomers, 28.4% of all homogenized monomers), of which roughly half (1521 monomers) are observed to also show intra-chromosomal homology (Supplemental Table 1-3, Supplemental Figure 1).

In line with previous studies [7-10], we observed increased local homogenization patterns directly adjacent to the centromeric gaps in the hg19 assembly, confirming the presence of 13 well-known, experimentally characterized higher-order repeat arrays, as well as two multimeric arrays with limited prior characterization found on chromosomes 4 and 11. Although the majority of type 1 monomers (3830, 86.9%) could be assigned to these 15 higher-order arrays, we also detected 577 monomers that appear to be involved in local sequence homogenization events within the genomic context of divergent monomeric alpha satellite, perhaps representing recent and limited homogenization events [11] (Supplemental Table 1). These regions could correspond to patterns that are polymorphic between different genomes, thus potentially confounding attempts to accurately map these refractory parts of the reference assembly.

While type 1 patterns have been well characterized previously, types 2 and 3 patterns appear to be novel, representing sequence exchanges between non-adjacent monomers, often spaced at considerable distances throughout the human genome. Such monomers tend to be arranged within regional domains, as 45.3% of all intra-chromosomal monomers share high homology with another monomer (or several other monomers) within a 500 kb window. There are, however, cases where monomers are homogenized at substantial distances, as we observe instances of monomer homology spanning eight centromere gaps (that is, from the p-arm side of the gap to the q-arm side) and/or between alpha satellite subsets that are located in non-centromeric positions on the chromosome arms [8]. These long distance patterns are particularly notable in the case of inter-chromosomal homology. For example, there are 860 assembled monomers (totaling ~183 kb) from the p-arm and q-arm of chromosome 9 that demonstrate high homology with 1,669 monomers from 14 different chromosomes. The majority of inter-chromosomal homogenized monomers appear to be interspersed within adjacent divergent monomers, possibly providing evidence for sites of long-range sequence conversion within centromeric DNA. Indeed, only 20.2% (625 monomers) of type 3 monomers and 5.7% (255 monomers) of type 2 monomers are observed to overlap with known segmental duplications in the human genome assembly. In total, we have identified 3696 pairwise intra- and inter-chromosomal long distance homogenization patterns.

Previous efforts to identify alpha satellite sequence features associated with centromere activity placed emphasis on both the presence of higher-order repeats (as reflected by many of the type 1 patterns in abundant homogenized arrays) and the presence of a centromere protein B binding site, the so-called "CENP-B box" [12-15]. Although 74.1% of all CENP-B boxes are found within homogenized higher-order repeats (Supplemental Table 4), we found 342 CENP-B boxes that are not associated with homogenized monomers, both within and outside of centromeric regions. It is not clear what role, if any, these CENP-B boxes play within the context of monomeric alpha satellite.

Supplemental References

1. Sullivan BA, Schwartz S: **Identification of centromeric antigens in dicentric Robertsonian translocations: CENP-C and CENP-E are necessary components of functional centromeres.** *Hum Mol Genet* 1995, **4:**2189-2197.
2. Harrington JJ, Van Bokkelen G, Mays RW, Gustashaw K, Willard HF: **Formation of de novo centromeres and construction of first-generation human artificial microchromosomes.** *Nat Genet* 1997, **15:**345-355.
3. Grimes BR, Rhoades AA, Willard HF: **Alpha-satellite DNA and vector composition influence rates of human artificial chromosome formation.** *Molecular therapy : the journal of the American Society of Gene Therapy* 2002, **5:**798-805.
4. Alexandrov IA, Medvedev LI, Mashkova TD, Kisselev LL, Romanova LY, Yurov YB: **Definition of a new alpha satellite suprachromosomal family characterized by monomeric organization.** *Nucleic Acids Res* 1993, **21:**2209-2215.
5. Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF: **Genomic and genetic definition of a functional human centromere.** *Science* 2001, **294:**109-115.
6. Shepelev VA, Alexandrov AA, Yurov YB, Alexandrov IA: **The evolutionary origin of man can be traced in the layers of defunct ancestral alpha satellites flanking the active centromeres of human chromosomes.** *PLoS Genet* 2009, **5:**e1000641.
7. Alexandrov IA, Kazakov AE, Tumeneva I, Shepelev V, Yurov Y: **Alpha-satellite DNA of primates: old and new families.** *Chromosoma* 2001, **110:**253-266.
8. Rudd MK, Willard HF: **Analysis of the centromeric regions of the human genome assembly.** *Trends Genet* 2004, **20:**529-533.
9. She X, Horvath JE, Jiang Z, Liu G, Furey TS, Christ L, Clark RA, Graves T, Gulden CL, Alkan C, et al: **The structure and evolution of centromeric transition regions within the human genome.** *Nature* 2004, **430:**857-864.
10. Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, Abrusan G: **Analysis of the largest tandemly repeated DNA families in the human genome.** *BMC Genomics* 2008, **9:**533.

11. Rudd MK, Wray GA, Willard HF: **The evolutionary dynamics of alpha-satellite.** *Genome Res* 2006, **16:**88-96.
12. Pluta AF, Saitoh N, Goldberg I, Earnshaw WC: **Identification of a subdomain of CENP-B that is necessary and sufficient for localization to the human centromere.** *J Cell Biol* 1992, **116:**1081-1093.
13. Muro Y, Masumoto H, Yoda K, Nozaki N, Ohashi M, Okazaki T: **Centromere protein B assembles human centromeric alpha-satellite DNA at the 17-bp sequence, CENP-B box.** *J Cell Biol* 1992, **116:**585-596.
14. Warburton PE, Waye JS, Willard HF: **Nonrandom localization of recombination events in human alpha satellite repeat unit variants: implications for higher-order structural characteristics within centromeric heterochromatin.** *Mol Cell Biol* 1993, **13:**6520-6529.
15. Masumoto H, Masukata H, Muro Y, Nozaki N, Okazaki T: **A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite.** *J Cell Biol* 1989, **109:**1963-1973.